Analysis of 106 kb of contiguous DNA sequence from the D genome of wheat reveals high gene density and a complex arrangement of genes related to disease resistance

Steven A. Brooks, Li Huang, Bikram S. Gill, and John P. Fellers

Abstract: Vast differences exist in genome sizes of higher plants; however, gene count remains relatively constant among species. Differences observed in DNA content can be attributed to retroelement amplification leading to genome expansion. Cytological and genetic studies have demonstrated that genes are clustered in islands rather than distributed at random in the genome. Analysis of gene islands within highly repetitive genomes of plants like wheat remains largely unstudied. The objective of our work was to sequence and characterize a contiguous DNA sequence from chromosome 1DS of *Aegilops tauschii*. An RFLP probe that maps to the *Lr21* region of 1DS was used to isolate a single BAC. The BAC was sequenced and is 106 kb in length. The contiguous DNA sequence contains a 46-kb retroelement-free gene island containing seven coding sequences. Within the gene island is a complex arrangement of resistance and defense response genes. Overall gene density in this BAC is 1 gene per 8.9 kb. This report demonstrates that wheat and its relatives do contain regions with gene densities similar to that of *Arabidopsis*.

Key words: resistance gene block, nucleotide-binding site, pathogenesis-related genes.

Résumé : Des différences considérables existent quant à la taille des génomes chez les plantes supérieures. Cependant, le nombre de gènes est relativement constant d'une espèce à l'autre. Les différences observées au niveau du contenu génomique en ADN peuvent être attribuées à l'amplification de rétroéléments, laquelle entraîne une expansion de la taille du génome. Des études cytologiques et génétiques ont montré que les gènes sont groupés en îlots plutôt que d'être distribués aléatoirement au sein du génome. Les îlots géniques au sein de génomes très répétitifs tels que celui du blé ont été peu étudiés. L'objectif de ce travail était de séquencer et de caractériser une séquence d'ADN contigu provenant du chromosome 1DS de l'*Aegilops tauschii*. Une sonde RFLP située au sein de la région *Lr21* de 1DS a été employée afin d'isoler un clone BAC. Ce clone a été séquencé et totalise 106 kb. Cette séquence d'ADN contient un îlot génique de 46 kb dépourvu de rétroéléments et compte sept régions codantes. Un arrangement complexe de gènes de résistance et de défense est observé au sein de l'îlot génique. La densité génique au sein de ce BAC est d'un gène à tous les 8,9 kb. Ce travail montre que le blé et les espèces apparentées contiennent des régions ayant des densités géniques semblables à celle observée chez *Arabidopsis*.

Mots clés : bloc de gènes de résistance, site de liaison de nucléotides, gènes de défense.

[Traduit par la Rédaction]

Introduction

The bread wheat (*Triticum aestivum* L.) genome (AABBDD) ($2n = 6x = 42$) is ~15 966 Mb (Arumuganathan and Earle 1991) and contains ~80% repetitive sequences (Smith and Flavell 1975). It is an allohexaploid species of three hybrid genomes designated A, B, and D. The ploidy level, size, and repetitive nature of the genome has impeded wheat genomics research. Reduction of this complexity is achieved when diploid genome donor species (A, B, or D) are analyzed as subgenomes of common wheat (Kam-Morgan et al. 1989; Stein et al. 2000). The D genome, derived from *Aegilops tauschii* Coss. (DD), is the smallest of all related genomes (~4024 Mb) (Arumuganathan and Earle 1991). Furthermore, *Ae. tauschii* is an important resource for agronomically important genes, as well as a resource for germplasm diversity in common wheat (Gill et al. 1991).

Cytological and genetic data have demonstrated the existence of "gene islands" distributed in the wheat genome (Werner et al. 1992; Gill et al. 1996; Faris et al. 2000; Sandhu et al. 2001). These gene-rich islands contain little or no repetitive DNA (Feuillet and Keller 1999), whereas neighboring gene-poor regions are highly repetitive. These observations support molecular evidence of genome expansion mechanisms through retroelement insertion and recombination events (Shirasu et al. 2000; Wicker et al. 2001).

Received 17 January 2002. Accepted 27 May 2002. Published on the NRC Research Press Web site at http://genome.nrc.ca on 20 September 2002.

Corresponding Editor: F. Belzile.

S.A. Brooks, L. Huang, and B.S. Gill. Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, U.S.A.
J.P. Fellers.[1] USDA–ARS, Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, U.S.A.

[1]Corresponding author (e-mail: jpf@alfalfa.ksu.edu).

 DOI: 10.1139/G02-049

Wheat genome analysis at the DNA sequence level for agronomically important loci is feasible if gene-rich islands are targeted for analysis.

Our goal is to characterize a gene-rich region by providing DNA-level analysis of a resistance-gene island from *Ae. tauschii* chromosome 1D. We sequenced 106 kb from a BAC library clone of *Ae. tauschii* (Moullet et al. 1999). The clone (M11) maps proximal to the *Lr21* leaf rust resistance locus on chromosome 1DS (Spielmeyer et al. 2000*a*) in a region known to possess resistance gene like sequences, including a marker that also cosegregates with *Lr21* (Spielmeyer et al. 2000*b*). Clone M11 contains a large retroelement-free gene island comprising >43% of the contig where high gene density is observed (1 gene per 6.6 kb). Within the island is a cluster of genes related to disease resistance. These genes display a new level of gene organization complexity for disease-resistance loci.

Materials and methods

Identification of low-copy BAC clones

A BAC library of *Ae. tauschii* accession AUS18913 (Moullet et al. 1999) was the source of clones for sequencing. Low copy sequence in the BAC clones was estimated by restriction digestion with *Bam*HI and with *Xba*I–*Xho*I in combination. Digested DNA was separated on 1% w/v agarose gels, stained with ethidium bromide, and documented by gel photography. DNA was transferred to nylon membrane and probed with ^{32}P-labeled wheat genomic DNA ('Chinese Spring'). Bands absent in exposed film versus the agarose gel were presumed to represent low copy sequences. Enzyme digestion, gel electrophoresis, Southern blotting, probe labeling, and hybridization were performed following the protocols described by Sharp et al. (1988).

BAC preparation and subcloning

Fifty micrograms of BAC M11 DNA was prepared using the QIAGEN® Large-Construct Kit (QIAGEN, Valencia, Calif.). Five micrograms of BAC DNA was sheared at 5 psi (1 psi = 6.894757 kPa) for 60 s using nebulizers (Invitrogen, Carlsbad, Calif.), resulting in DNA fragments of 1–5 kb in size. Sheared DNA was blunt-end repaired and dephosphorylated without size selection (Invitrogen). Thirty nanograms of sheared BAC DNA was used as an insert for ligation into the pCR®4Blunt-TOPO® vector and transformed into TOP10 One Shot® chemically competent *Escherichia coli* cells (F$^-$ *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*Z ΔM15 Δ*lac*74 *rec*A1 *deo*R *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L (Str^R) *end*A1 *nup*G) (Invitrogen). There were 768 single colonies selected and arrayed into 96-well plates to produce liquid cultures for plasmid purification. Plasmids were isolated with a QIAGEN® BioRobot™ 3000, using QIAprep® 96 Turbo BioRobot™ Kits (QIAGEN).

Sequence analysis and contig assembly

Subclone plasmids were sequenced directly with 5 pmol of primer (T3 and T7) in 10-μL reactions using ABI Prism® BigDye™ Terminator Ready Reaction Cycle Sequencing kits (Applied Biosystems, Foster City, Calif.). Completed sequencing reactions were run on an ABI Prism® 3700 DNA Analyzer, base quality scores called by phred version 0.990722.f (Ewing et al. 1998; Ewing and Green 1998), and contigs assembled using phrap version 0.990319 (http://www.phrap.org). Consed version 11.0 (Gordon et al. 1998) was used to edit contigs and design custom primers to complete the sequence of clones spanning gaps. Manual alignment of sequences from clones spanning gaps to assembled contigs was performed with AssemblyLIGN™ version 1.0.9c (Oxford Molecular Ltd., Madison, Wis.).

Sequence annotation and CDS identification

FASTA files of assembled contigs were exported from Consed for coding sequence (CDS) identification. GENSCAN 1.0 (http://genes.mit.edu/GENSCAN.html) was used to predict CDSs with maize.smat as the parameter matrix. In addition, FGENESH 1.1 (http://www.softberry.com) was used for CDS prediction with monocot genomic DNA parameters. Both programs predict promoter regions and polyadenylation signals for each CDS. CDSs identified by both programs were used for subsequent analysis. Putative polypeptide sequences were defined by results of BLASTp and BLASTx searches against the nonredundant (nr) NCBI database (Altschul et al. 1997; http://www.ncbi.nlm.nih.gov/BLAST/). Perfect microsatellite repeats were detected using the simple sequence repeat identification tool (SSRIT) (http://www.gramene.org/gramene/searches/ssrtool/).

Results

BAC clone M11 was determined to be rich in low copy sequences by Southern hybridization with ^{32}P-labeled wheat genomic DNA. Comparison of Southern blots to ethidium bromide stained agarose gel photographs revealed that 9 out of 21 bands did not hybridize in the *Bam*HI digestion, and 6 out of 10 bands did not hybridize in the *Xba*I–*Xho*I double digestion. These results indicate that M11 is rich in low-copy sequences (data not shown).

M11 was previously mapped to *Ae. tauschii* chromosome 1DS, and contains two distinct R gene like sequences (Spielmeyer et al. 2000*a*). One sequence is detected by genomic DNA clone KSUD14, which was derived from a *Pst*I-digested genomic DNA library of *Ae. tauschii* accession TA1691 (Gill et al. 1991) and is 1.36 kb in length. KSUD14 contains a putative open reading frame fragment of 194 amino acids (aa), including two nucleotide binding site (NBS) like motifs (kinase2a and kinase3 domains of NBS – leucine-rich repeat (LRR) class R genes). Amino acid sequences from KSUD14 show high similarity to the putative cereal cyst nematode (CCN) resistance gene *Cre3* from wheat (Lagudah et al. 1997). *XksuD14* was 0.2-cM proximal to *Lr21* in a cross with WGRC2 and cosegregated with *Lr21* in a cross with WGRC7 (Huang and Gill 2001).

Sequence analysis

A contiguous DNA sequence of 106 618 bp was obtained from BAC M11. Sequencing revealed a 44.45% GC content (GenBank accession No. AF446141). Complete DNA sequence analysis of M11 reveals high gene density and an island of genes involved in host plant resistance to disease (Fig. 1). GENSCAN predicted 22 putative coding sequences (CDSs) in the contig, and FGENESH predicted 17. Analysis of the predicted CDSs revealed similar results from each

Fig. 1. M11 contig map. One hundred six kilobases of contiguous DNA sequence from *Ae. tauschii* chromosome 1DS. The positions of predicted CDSs are indicated by number above the contig. Red boxes are retroelement-like sequences, green boxes are disease- and (or) defense-related sequences, yellow boxes indicate strong alignments to hypothetical proteins, and purple boxes indicate sequences with poor or no database alignment. The tandem repeat sequence is shown with black cross hatching. The positions of SSRs are indicated by number below the contig.

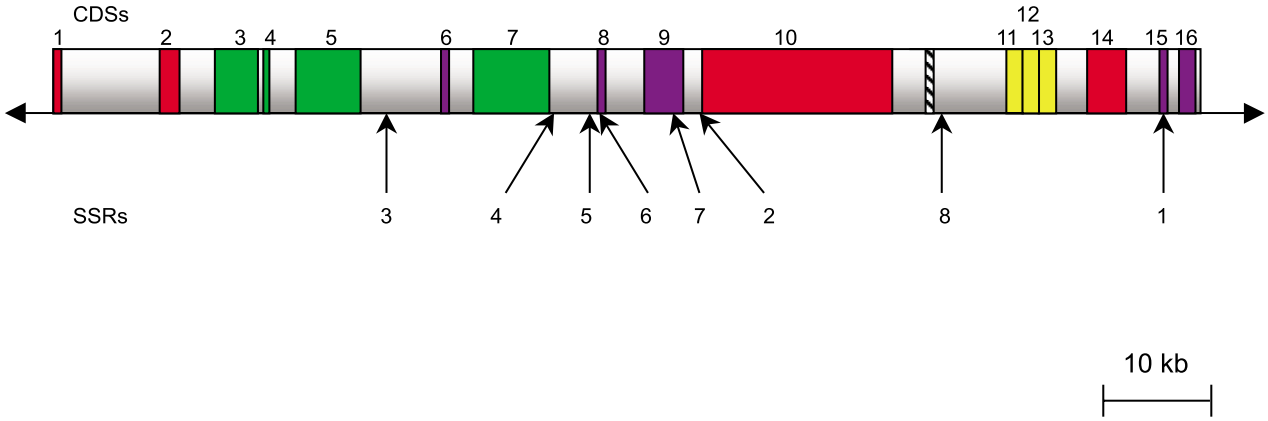


Table 1. Non-retroelement coding regions in BAC M11 predicted by GENSCAN and FGENESH.

CDS	Start	Stop	O	BLASTx results	GenBank accession No.	Score	*E* value
3	15 577	19 353	+	Stripe rust resistance protein Yr10 (wheat)	AF149112	832	0
4	20 866	21 103	+	*B. napus* receptor protein kinase PERK1	AY028699	39	0.01*
5	23 271	29 048	–	Probable cyst nematode R-gene (wheat)	AF052641	649	0
6	35 813	36 623	+	poor alignment	—	—	—
7	39 402	46 082	+	*N*-hydroxycinnamoyl / benzoyltransferase-like protein (*A. thaliana*)	AB008264	262	3.00×10^{-68}
8	50 631	51 656	+	No significant similarity found	—	—	—
9	55 803	59 071	+	poor alignment	—	—	—
11	88 386	91 045	–	Hypothetical protein F17K2.16 (*A. thaliana*)	T00876	229	1×10^{-58}
12	91 214	92 529	–	Hypothetical protein (*Oryza sativa*)	AP002524	48	3×10^{-4}
13	92 762	94 353	+	Alliin lyase homolog F22K18.130 (*A. thaliana*)	T05567	321	2×10^{-86}
15	103 358	103 863	–	poor alignment	—	—	—
16	105 036	106 135	+	poor alignment	—	—	—

Note: CDS position is indicated by number of the first nucleotide of the start codon and the last nucleotide of the stop codon. The direction of transcription (O) and BLASTx results of predicted nucleotide sequence are indicated. GenBank accession numbers, scores (in bits), and *E* values are provided for each BLAST alignment. Poor alignments have an *E* value > 0.0001. Asterisk indicates poor probability of alignment.

program. A consensus was taken and 16 individual coding regions were identified (Fig. 1 and supplemental data[2]). BLASTx searches performed with sequences between the predicted CDSs (intergenic) show no significant database alignments, indicating the utility of GENSCAN and FGENESH for predicting coding sequences. Twelve of the predicted CDSs are non-repetitive putative gene-coding regions, resulting in an overall gene density of one gene per 8884 bp (Table 1). The remaining four predicted CDSs were classified as retroelement-like based on BLASTx results (Table 2).

CDS identity was predicted by BLASTp searches of GENSCAN-predicted polypeptide sequences against the NCBI nr database. To support the predicted identity, BLASTx searches of the same database were performed using the nucleotide sequences corresponding to the predicted CDSs (Table 1). BLASTx results for 12 CDSs (75%) supported the BLASTp results. BLASTx results of the remaining four CDSs (CDSs 2, 4, 5, 16) had different identities than those predicted by BLASTp. In these cases, GENSCAN predicted the gene positions appropriately (see above); however, the putative polypeptide sequence is better defined by BLASTx translation and (or) database alignments.

To support CDS prediction, alignments were made to wheat and barley (*Hordeum vulgare*) EST databases using the BLASTn algorithm (WUBLAST 2.0); The Institute for Genomic Research (TIGR) gene indices TaGI and HvGI, respectively (http://tigrblast.tigr.org/tgi). Of the 12 non-retroelement-like CDSs, 9 produced significant alignments in both EST databases (CDSs 3, 4, 5, 6, 7, 11, 13, 15, and

[2] Supplementary material may be purchased from the Depository of Unpublished Data, Document Delivery, CISTI, National Research Council Canada, Ottawa, ON K1A 0S2, Canada. For information on ordering material electronically go to http://www.nrc.ca/cisti/irm/unpub_e.shtml.

Table 2. Retroelement-like regions in BAC M11.

CDS	Start	Stop	Type	BLASTx results	GenBank accession No.	Score	*E* value
1	1	422	Ty1-*copia*	Putative *copia*-like retrotransposon polyprotein (*Oryza sativa*)	AC073166	181	2×10^{-45}
2	10 040	11 032	Ty1-*copia*	Putative gag-pol polyprotein (*Oryza sativa*)	AC079037	447	10^{-124}
10	60 965	78 700	complex	—	—	—	—
14	96 355	99 711	Ty3-*gypsy*	Putative polyprotein (*Zea mays*)	AAL75999	808	0.0

Note: Regions are inclusive of indicated contig positions. The type of element is indicated along with the BLASTx results of the corresponding nucleotide sequence. GenBank accession numbers, scores (in bits), and *E* values are provided for each BLAST alignment.

Table 3. Predicted coding sequences aligned to wheat and barley ESTs using the BLASTn algorithm (WUBLAST 2.0).

CDS	Wheat EST	Score	*E* value	Barley EST	Score	*E* value
3	BE585677	930	4.1×10^{-36}	TC6133	1173	1.4×10^{-90}
4	BG909158	212	5.5×10^{-5}	BI951125	222	0.0001
5	BE498831	1380	2.6×10^{-56}	AU090217	1000	1.4×10^{-38}
6	TC14726	1856	1.3×10^{-79}	BI949148	344	4.0×10^{-8}
7	BE446216	756	9.6×10^{-28}	BI952010	844	1.1×10^{-31}
8	none	—	—	BF629167	285	4.3×10^{-5}
9	poor	—	—	BF260452	384	8.1×10^{-9}
11	BG904777	721	1.0×10^{-26}	BI960579	1130	1.9×10^{-45}
12	poor	—	—	Poor	—	—
13	BE423530	338	9.0×10^{-10}	AL508912	1184	8.9×10^{-78}
15	BG313070	442	1.7×10^{-15}	BI956794	250	2.4×10^{-6}
16	BF428736	603	6.8×10^{-23}	BG343104	629	1.5×10^{-35}

Note: TIGR gene indices TaGI and HvGI respectively (http://tigrblast.tigr.org/tgi). Scores (in bits) for BLAST results and *E* values are provided for each EST. Poor alignments have an *E* value > 0.0001.

16), 2 in barley only (CDSs 8 and 9), and one (CDS12) did not show any alignments to either database (Table 3).

Resistance-associated gene island

M11 contains an island of genes related to host plant resistance to disease. This island begins 2138 bp downstream of a Ty1-*copia* retroelement-like sequence (CDS2). The region is 34 397 bp in length, starting with the promoter for CDS3 (position 13169) and ending with the polyadenylation signal for CDS7 (position 47566). The region contains five putative genes, four of which show similarity to known genes involved in disease resistance, and one that has no known function (CDS6).

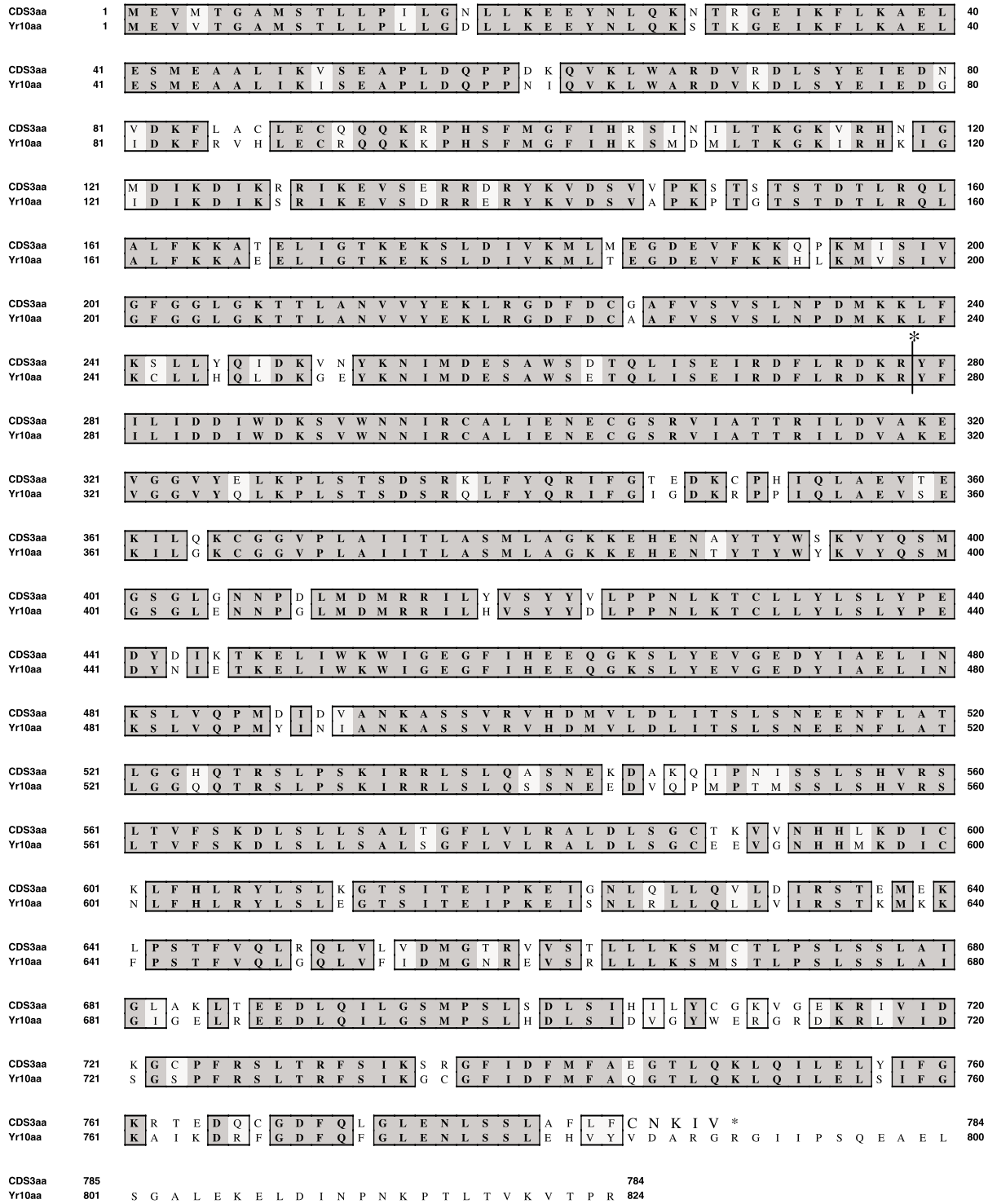
CDS3 is the first predicted gene in the resistance-gene island and is in the plus orientation. The GENSCAN-predicted polypeptide is 299 amino acids long, showing homology to the putative stripe rust resistance protein *Yr10* of wheat (GenBank accession No. AF149112) up to amino acid 267 of the predicted protein. This is at the predicted 5′ end of the *Yr10* intron, after which alignment to *Yr10* stops (not shown). However, BLASTx results of the nucleotide sequence for CDS3 gives a gapped alignment to *Yr10*, indicating an alternate intron position corresponding to the intron predicted for *Yr10*. By combining BLASTx-predicted gapped polypeptide sequences into a single protein sequence, complete alignment of CDS3 to *Yr10* is achieved (Fig. 2). The 5′ end of exon 1 for this 784-aa sequence is at the same position predicted by GENSCAN (position 15 577). The putative intron falls between amino acids 278 and 279 of the kinase 2a domain, inclusive of nucleotide positions 16 411 through 17 564 of the contig. The 3′ end of the terminal exon is at position 19 082 of the contig. The amino acid sequence predicted from BLASTx alignment is followed by an additional five amino acids with no similarity to the same positions in *Yr10*, before an in-frame termination codon (CNKIV*, Fig. 2).

CDS3 and *Yr10* possess all five conserved domains characteristic of NBS–LRR-class resistance proteins described by Grant et al. (1995). Table 4*a* shows the position and amino acid sequences for the P-loop, kinase 2a, kinase 3a, conserved domain 2, and conserved domain 3. There is 100% conservation of these motifs with *Yr10* at the amino acid level. There is also perfect nucleotide conservation within all motifs except the P-loop, where a single nucleotide difference is found at position 16 197.

CDS4 lies 1514 bp downstream of CDS3 and is in the same (plus) orientation. GENSCAN predicted a 62-aa polypeptide with no homology to known protein sequences. However, BLASTx searches of the nucleotide sequence of CDS4 shows poor alignment to the *Brassica napus* receptor protein kinase *PERK1* (Table 1). BLASTx results show similarity in a 42-aa sequence from the nucleotide positions 20 952 through 21 077, lying within the predicted CDS, in a +2 reading frame relative to the GENSCAN-predicted reading frame. Alignment of this sequence occurs within the predicted kinase domain of *PERK1* (not shown). Although alignment at the amino acid level is poor, significant alignments were made to wheat and barley ESTs (Table 3).

CDS5 is the second NBS–LRR-class gene in the resistance-gene island, lying 2169 bp downstream of CDS4 in an opposing (minus) orientation. GENSCAN predicted a 640-aa polypeptide sharing alignment with rust resistance protein

Fig. 2. ClustalW amino acid alignments of *CDS3aa* and *Yr10aa* (NCBI protein identification number AAG42167.1) using MacVector™ 6.5.3 (784 and 824 amino acids, respectively). Dark shading indicates amino acid identity, light shading indicates amino acid similarity, and no shading indicates no amino acid similarity. Asterisk indicates position of the predicted intron.



Rp1-dp2 of *Zea mays*. However, a more significant alignment is achieved with BLASTx searches of the nucleotide sequence for CDS5. BLASTx results show alignment to the cereal cyst nematode resistance gene candidate(*Cre3*) of wheat (GenBank accession No. AF052641; Table 1). The predicted 795 aa sequence corresponds to nucleotide positions 26864 through 24479 of the contig followed by an in frame stop codon. The amino acid sequence shows ungapped alignment to the partial coding sequence (mRNA) of *Cre3*, indicating a lack of intron(s) in this sequence (Fig. 3).

The amino acid sequence predicted by BLASTx for CDS5 contains all five conserved domains characteristic of NBS–LRR-class resistance proteins described by Grant et al.

Table 4. Five conserved domains of NBS–LRR-class resistance proteins as described by Grant et al. (1995).

(*a*) NBS conserved domains of CDS3.			
Domain	Motifs	Amino acid position	Nucleotide position
P-loop	GFGGLGKTT	201–209	16177–16203
Kinase 2a	KRYFILIDDI	277–286	16405–17588 *
Kinase 3a	GSRVIATTRILDV	305–317	17643–17681
Conserved domain 2	CGGVPLAIITLAS	366–378	17826–17864
Conserved domain 3	LKTCLLY	428–434	18012–18032
(*b*) NBS conserved domains of CDS5.			
Domain	Motifs	Amino acid position	Nucleotide position
P-loop	GVAGSGKTT	89–97	26599–26573
Kinase 2a	KRFLLILDDL	166–175	26368–26339
Kinase 3a	GSKILVTARTKEA	198–210	26272–26234
Conserved domain 2	LHGSPIAAVTVAG	260–272	26086–26048
Conserved domain 3	IRRCFEF	310–316	25936–25916

Note: The domain, motif, amino acid, and nucleotide positions are indicated.
*Position of an intron within the kinase 2a domain.

(1995). Table 4*b* shows the position and amino acid sequences for the P-loop, kinase 2a, kinase 3a, and conserved domains 2 and 3. Conservation of these motifs at the amino acid level is varied in CDS5 and *Cre3*. Seventy-eight percent of amino acid residues are conserved in the P-loop, 80% in kinase 2a, 77% in kinase 3a, 69% in conserved domain 2, and 57% in conserved domain 3.

CDS6 lies 6766 bp downstream of CDS5 and is in the plus orientation. The GENSCAN-predicted 114-aa polypeptide sequence shows no significant homology to known protein sequences. Furthermore, BLASTx searches of the nucleotide sequence of CDS6 show no significant database alignments. However, support for the CDS prediction is provided by significant EST hits from wheat and barley (Table 3).

CDS7 lies 2780 bp downstream of CDS6 in the plus orientation, and is the last CDS in the resistance-gene island. The GENSCAN-predicted polypeptide is 558 aa, showing homology to the *Arabidopsis N*-hydroxycinnamoyl / benzoyltransferase (HCBT) like protein, a putative defense-related protein. BLASTx translation alignments of the CDS7 nucleotide sequence confirm the BLASTp results, and has complete in-frame amino acid alignment with the GENSCAN-predicted polypeptide sequence (not shown). Furthermore, the BLASTx translation has a 459-aa sequence alignment with 35% amino identities and 53% amino acid similarities to the *Arabidopsis* protein.

The resistance-gene island is followed immediately by two genes with no known function, positioned 4550 bp downstream of CDS7. BLASTp and BLASTx searches of nucleotide and protein sequences corresponding to CDSs 8 and 9 produced no significant database alignment, but did produce EST hits in barley (Tables 1 and 3). Taken together, the resistance-gene island and CDSs 8 and 9 total 46 544 bp in length, comprising 43% of the contig, and contain no retroelement-like sequences.

Remaining genes in BAC M11

The remainder of the contig contains five CDSs separated from the gene island by a highly repetitive region of multiple retroelement insertions (CDS10). CDSs 11 and 13 produced significant database alignments with both BLASTp and BLASTx algorithms. In both cases, BLASTp and BLASTx searches identified the same protein sequence from the database. CDS11 aligns to a hypothetical protein from *Arabidopsis* (259-aa BLASTx alignment) with 47% amino acid identity and 62% amino acid similarity. CDS13 aligns to an alliin lyase homolog from *Arabidopsis* (427-aa BLASTx alignment) with 42% amino acid identity and 54% amino acid similarity. CDS12 did not produce any significant alignments using BLASTp. However, CDS12 shows similarity to a hypothetical protein from *O. sativa* (71-aa BLASTx alignment) with 36% amino acid identity and 21% amino acid similarity. CDSs 15 and 16 produced no significant database alignments with BLASTp or BLASTx algorithms; however, both produced EST hits in wheat and barley (Tables 1 and 3).

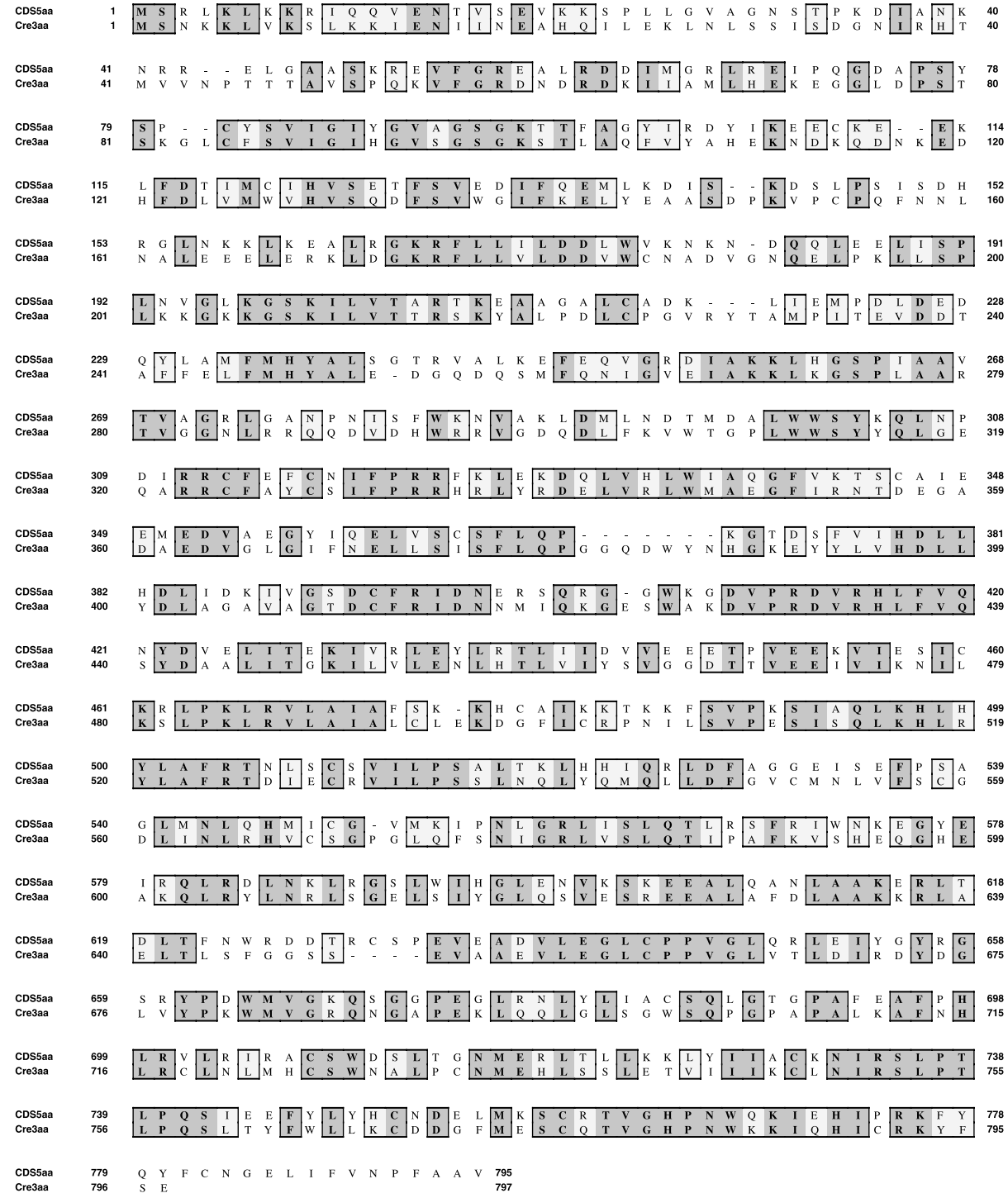
Mobile elements, repetitive sequences, and microsatellites

Retroelement-like sequences were predicted by GENSCAN and FGENESH in M11 as CDSs and were identified by BLASTx results (CDSs 1, 2, 10, and 14) (Table 2). CDSs 1 and 2 are long terminal repeat (LTR) type retrotransposons of the Ty1-*copia* subgroup. CDS14 is also an LTR-type retroelement, but is of the Ty3-*gypsy* subgroup. CDS10 represents a highly repetitive region composed of multiple insertions of retroelements spanning 17 736 bp of the contig. The region spans positions 60 965 – 78 700 of the contig 1895 bp downstream of CDS9 and 9687 bp upstream of CDS11.

A highly repetitive non-transposable sequence of 890 bp was identified spanning positions 81 907 – 82 796 of the contig. The sequence is highly homologous to a tandem repeat sequence from *Aegilops squarrosa* (syn. *tauschii*) (GenBank accession No. D30736, 1558 bit-score, *E* value = 0.0 by BLASTn search of NCBI nr database), positioned 3208 bp downstream of the CDS10 region and 5591 bp upstream of CDS11 (Fig. 1).

Twenty-three perfect simple sequence epeats (SSRs) were identified in BAC M11. Dimers of less than five repeats in length were omitted, resulting in eight SSRs of interest (Table 5). The physical positions of the SSRs in the contig are

Fig. 3. ClustalW amino acid alignments of *CDS5aa* and *Cre3aa* (NCBI protein identification No. AAC05834.1) using MacVector™ 6.5.3 (795 and 797 amino acids, respectively). Dark shading indicates amino acid identity, light shading indicates amino acid similarity, and no shading indicates no amino acid similarity.



indicated in Fig. 1. Pairs of SSRs identified in close proximity were combined as imperfect repeats (SSRs 1, 3, and 6). SSR 1 is an AT dimer motif of 28 repeats interrupted by an AAA trimer between the 10th and 11th repeats. SSR 3 is a TA dimer motif of 19 repeats followed an AC dimer motif of 6 repeats. The two motifs share a common A residue in the terminal TA dimer and the initial AC dimer (TAC). SSR 6 is a TG dimer motif of 10 repeats interrupted by TACAC between the 5th and 6th repeats. SSRs 3 and 4 are of interest owing to their close proximity to genes potentially involved in disease resistance. SSR 3 is 1660 bp upstream of the initial exon of an NBS–LRR-class resistance-gene analog, and 144 bp upstream of its predicted promoter (CDS5). SSR 4 is 1104 bp from the 3′ end of the terminal exon of a putative

Table 5. M11 simple sequence repeats.

Sequence	Motif	No. of repeats	SSR start	SSR end
SSR 3	$TA_{19}AC_6$	25	30709	30760
SSR 4	AG	37	47187	47260
SSR 5	GA	31	50218	50279
SSR 6	$TG_5(TACAC)TG_5$	10	51314	51338
SSR 7	ATG	8	58046	58069
SSR 2	AG	33	60541	60606
SSR 8	AGA	5	83823	83837
SSR 1	$AT_{10}(AAA)AT_{18}$	28	102976	103034

Note: SSRIT parameters were set to find perfect repeats up to one decamer, with a minimum repeat number of five. The repeat motif, number of repeats, and SSR positions are indicated. Dimers of five repeats or less have been omitted.

pathogenesis related (PR) gene (CDS7), and 300 bp upstream of its poly(A) signal, located in the putative 3′ untranscribed region (UTR) of the gene.

Discussion

BAC clone M11 is located in an agronomically important region of the *Aegilops tauschii* genome. M11 is positioned genetically distal to a seed storage protein locus (*Gli1–Glu-3*), and proximal to the *Lr21* leaf rust resistance locus on the short arm of chromosome 1 (Spielmeyer et al. 2000*a*). M11 has also been shown to be rich in low-copy sequences and to contain resistance gene like sequences, thus making it an ideal clone for directed genome analysis of a gene-rich region.

The overall gene density in M11 is one gene per 8.9 kb, the highest reported thus far in genomic analyses of grasses. In barley, two independent studies reported an overall gene density of one gene per 20 kb (Shirasu et al. 2000; Panstruga et al. 1998) on chromosomes 2HL and 4HL, respectively. A gene density of one gene per 42 kb was also reported for chromosome 1AS of *Triticum monococcum* L. (Wicker et al. 2001). The gene density observed within the retroelement-free gene island of M11 is one gene per 6.6 kb. This is the largest gene island sequence reported in a grass genome, with 7 CDSs located over 46 544 bp. This observation is comparable with the density observed in a barley gene island, where three genes were clustered in an 18-kb region (Shirasu et al. 2000).

At the 5′ end of the gene-rich region is an island of genes related to disease resistance. The complex arrangement of genes at this locus represents a new structure for resistance gene loci and has not been previously reported. The maize *Rp1-D* rust resistance locus represents a tandem arrangement of NBS–LRR-class resistance genes, where the structure of the locus provides a means for evolution of new rust resistance genes (Collins et al. 1999). Indeed, more complex arrangements exist for resistance loci in plants. In tomato (*Lycopersicon esculentum*), the *Pto* locus confers resistance to the bacterial pathogen *Pseudomonas syringae* pv. *tomato* (Pst). The locus contains at least six gene family members that show similarity to serine (threonine) protein kinases (Martin et al. 1993). Located within this gene cluster is *Prf*, a leucine zipper (LZ) – NBS – LRR class resistance gene (Salmeron et al. 1996). Both *Pto* and *Prf* are necessary for resistance to Pst strains expressing *avrPto*, indicating a clustering of different genes involved in the same resistance pathway. Another complex arrangement of genes is found at the *Lr10* leaf rust resistance locus in wheat. A receptor kinase like gene, *Lrk10*, has been mapped to the locus on chromosome 1AS, and is tightly linked to the *Lr10* resistance gene (Feuillet et al. 1997). Furthermore, an additional receptor kinase like gene (*Tak10*), and an LZ–NBS–LRR-class pseudogene (*ΨLrr10*) are found at the same locus (Feuillet and Keller 1999). Homologs of these gene classes are also found at orthologous loci in barley.

The presence of multigene families at disease-resistance loci in plants, coupled with evidence for more than one gene per locus being necessary for conferring resistance, raises the question that selection pressure may maintain multigene families as a single functional unit. The resistance gene island in M11 contains four unique genes, each belonging to different classes of disease- and (or) defense-related genes. There are two members of NBS–LRR-class resistance genes, differing in the N-terminal protein motifs. CDS3 (*Yr10* homolog) contains an LZ motif, whereas CDS5 (*Cre3* homolog) lacks additional N-terminal motifs. The conservation of the amino acid sequence of these genes with their corresponding homologs is markedly different. It is interesting that CDS3 and *Yr10* show such extensive homology when the two genes are derived from separate genomes (1DS and 1BS, respectively) that shared a common ancestor 2–4 million years ago (B.S. Gill[3]).

The conservation of sequence in CDS3 with *Yr10* is indicative of selection pressure for conserved function. Because *Yr10* confers resistance to stripe rust of wheat (*Puccinia striiformis* Westend. f.sp. *tritici*) (McIntosh et al. 1995), it is probable that CDS3 may confer resistance to a similar pathogen. Given that an RFLP marker (*rgaYr10b*) cosegregates with *Lr21* and identifies M11 (*rgaYr10a* = CDS3) in a region proximal to *Lr21* (Spielmeyer et al. 2000*b*), it is possible that CDS3 may confer resistance to an unknown pathotype of leaf or stripe rust.

CDS4 shows poor homology with a receptor kinase like gene from *Brassica napus*. It is possible that alignment occurred by chance; however, significant EST alignments in both wheat and barley provide support for a coding sequence. This may represent a degenerated or highly diver-

[3] B.S. Gill. Manuscript in preparation.

gent receptor kinase gene, because no alignments were made to receptor kinases from wheat and barley at the amino acid level. This sequence resides between the two NBS–LRR-class genes, reminiscent of the proximity of different gene classes observed at the *Lr10* locus in wheat (Feuillet and Keller 1999) and the *Pto* locus in tomato (Martin et al. 1993; Salmeron et al. 1996). Furthermore, a consistency is shown in the type of NBS–LRR gene found near kinase-like genes (serine (threonine) protein kinase and receptor kinase; *Pto* and *Lrk10*, respectively). CDS3, *Prf*, and Ψ*Lrr10* are all members of the LZ–NBS–LRR class of resistance genes, and are all tightly linked to kinase-like resistance genes. This consistency is interesting because the leucine zipper is predicted to be involved in protein–protein interactions, and it has been shown for *Pto* and *Prf* that both are active in the same resistance pathway.

The most unique feature of this locus is the tight linkage of a defense-related gene to disease-resistance gene analogs. CDS7 shows significant homology to an HCBT-like protein from *Arabidopsis*. In carnation (*Dianthus caryophyllus* L.) HCBT proteins catalyze the biosynthesis of unique dianthramide phytoalexins (Yang et al. 1998). HCBT transcription is induced rapidly and transiently by pathogen elicitors, and dianthramide accumulation is essential for resistance to fungal pathogens (Yang et al. 1997). To our knowledge, this is the first HCBT-like gene cloned from a wheat subgenome, and the first time such a gene has been shown in tight linkage with disease-resistance gene analogs.

Previous genomic analyses in grasses at the DNA sequence level have described retroelement insertion events in detail (Panstruga et al. 1998; Shirasu et al. 2000; Wicker et al. 2001). Transposable elements were identified flanking the 46-kb gene island and interspersed within a group of genes at the right end of the contig (Fig. 1). CDS 1 is a retroelement fragment interrupted at the left cloning site. CDSs 2 and 14 appear to be incomplete elements of single integration events. The CDS 10 region represents a complex of multiple retroelement insertions in a 17-kb region. The arrangement of elements is highly complex, reminiscent of the multiple insertions observed at the *Lr10* locus of *Triticum monococcum* L. (Wicker et al. 2001). Although outside the scope of the present work, it is interesting that many integrations have occurred in this small region.

SSRs identified in M11 are valuable markers for future analysis of homologous regions in wheat and *Ae. tauschii*. Six of the SSRs identified are located within the 46-kb gene island (Fig. 1). These markers are currently being used for genetic mapping to orient M11 on 1DS and to provide an assessment of physical : genetic distance in this region.

The present work provides molecular evidence in support of gene-rich islands containing no repetitive DNA (retroelements) in the wheat genome. Analysis of the resistance-gene island in M11 reveals a new level of complexity not yet reported in plants, where NBS–LRR-class R genes are tightly clustered with other defense-related genes. This data sheds new light on the structure of resistance gene loci, and provides evidence for a higher level of genetic complexity associated with host plant resistance to disease. Additional research at the DNA sequence level is necessary to broaden our vision of disease resistance loci. This work will provide the insight necessary to design functional resistance gene pyramids for durable or multiple resistances. Work is underway to characterize the expression of genes associated with disease resistance by RT–PCR, and to determine the level of sequence diversity within the resistance gene island in multiple accessions of *Ae. tauschii* and wheat. Our reverse genetics approach will provide insight into the function and conservation of these genes. Furthermore, mutagenesis and transgenic studies will lead to the final goal of phenotypic assessment of gene function.

Acknowledgements

We extend our thanks to the following people for their contributions to this project: Dr. Evans Lagudah for kindly supplying BAC clones for analysis and sequencing; Angie Matthews for sequencing and sequence analysis; Tawanna Gayle Ross-Vardeman for technical help and plasmid preparations; Dr. Kristi Hill-Ambroz for microsatellite analysis; and Marie Herbel for technical support. This manuscript is submitted as Journal number 02-190-J of the Kansas Agricultural Experiment Station. Mention of a trademark of a proprietary product does not constitute a guarantee of warranty of the product by the United States Department of Agriculture, and does not imply its approval to the exclusion of other products that may also be suitable.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**: 3389–3402.

Arumuganathan, K., and Earle, E.D. 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. **9**: 208–218.

Collins, N., Drake, J., Ayliffe, M., Sun, Q., Ellis, J., Hulbert, S., and Pryor, T. 1999. Molecular characterization of the maize *Rp1-D* rust resistance haplotype and its mutants. Plant Cell, **11**: 1365–1376.

Ewing, B., and Green, P. 1998. Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. Genome Res. **8**: 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using *Phred*. I. Accuracy assessment. Genome Res. **8**: 175–185.

Faris, J.D., Haen, K.M., and Gill, B.S. 2000. Saturation mapping of a gene-rich recombination hot spot region in wheat. Genetics, **154**: 823–835.

Feuillet, C., and Keller, B. 1999. High gene density is conserved at syntenic loci of small and large grass genomes. Proc. Natl. Acad. Sci. U.S.A. **96**: 8265–8270.

Feuillet, C., Schachermayr, G., and Keller, B. 1997. Molecular cloning of a new receptor-like kinase gene encoded at the *Lr10* disease resistance gene locus of wheat. Plant J. **11**: 45–52.

Gill, K.S., Lubbers, E.L., Gill, B.S., Raupp, W.J., and Cox, T.S. 1991. A genetic linkage map of *Triticum tauschii* (DD) and its relationship to the D genome of bread wheat (AABBDD). Genome, **34**: 362–374.

Gill, K.S., Gill, B.S., Endo, T.R., and Taylor, T. 1996. Identification and high-density mapping of gene-rich regions in chromosome group *1* of wheat. Genetics, **144**: 1883–1891.

Gordon, D., Abajian, C., and Green, P. 1998. *Consed*: a graphical tool for sequence finishing. Genome Res. **8**: 195–202.

Grant, M.R., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., Innes, R.W., and Dangl, J.L. 1995. Structure of the *Arabidopsis RPM1* gene enabling dual specificity disease resistance. Science (Washington, D.C.), **269**: 843–846.

Huang, L., and Gill, B.S. 2001. An RGA – like marker detects all known *Lr21* leaf rust resistance gene family members in *Aegilops tauschii* and wheat. Theor. Appl. Genet. **103**: 1007–1013.

Kam-Morgan, L.N.W., Gill, B.S., and Muthukrishnan, S. 1989. DNA restriction fragment length polymorphisms: a strategy for genetic mapping of D genome of wheat. Genome, **32**: 724–732.

Lagudah, E.S., Moullet, O., and Appels, R. 1997. Map-based cloning of a gene sequence encoding a nucleotide-binding domain and a leucine-rich region at the *Cre3* nematode resistance locus of wheat. Genome, **40**: 659–665.

Martin, G.B., Brommonschenkel, S.H., Chunwongse, J., Frary, A., Ganal, M.W., Spivey, R., Wu, T., Earle, E.D., and Tanksley, S.D. 1993. Map-based cloning of a protein kinase gene conferring disease resistance in tomato. Science (Washington, D.C.), **262**: 1432–1436.

McIntosh, R.A., Welling, C.R., and Park, R.F. 1995. Wheat rusts: an atlas of resistance genes. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Moullet, O., Zhang, H.B., and Lagudah, E.S. 1999. Construction and characterisation of a large DNA insert library from the D genome of wheat. Theor. Appl. Genet. **99**: 305–313.

Panstruga, R., Büschges, R., Piffanelli, P., and Schulze-Lefert, P. 1998. A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. Nucleic Acids Res. **26**: 1056–1062.

Salmeron, J.M., Oldroyd, G.E.D., Rommens, C.M.T., Scofield, S.R., Kim, H., Lavelle, D.T., Dahlbeck, D., and Staskawicz, B.J. 1996. Tomato *Prf* is a member of the leucine-rich repeat class of plant disease resistance genes and lies embedded within the *Pto* kinase gene cluster. Cell, **86**: 123–133.

Sandhu, D., Champoux, J.A., Bondareva, S.N., and Gill, K.S. 2001. Identification and physical location of useful genes and markers to a major gene-rich region on wheat group *1S* chromosomes. Genetics, **157**: 1735–1747.

Sharp, P.J., Chao, S., Desai, S., and Gale, M.D. 1988. The isolation, characterization and application in the Triticeae of a set of wheat RFLP probes identifying each homoeologous chromosome arm. Theor. Appl. Genet. **78**: 342–348.

Shirasu, K., Schulman, A.H., Lahaye, T., and Schulze-Lefert, P. 2000. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res. **10**: 908–915.

Smith, D.B., and Flavell, R.B. 1975. Characterisation of the wheat genome by renaturation kinetics. Chromosoma, **50**: 223–242.

Spielmeyer, W., Moullet, O., Laroche, A., and Lagudah, E.S. 2000*a*. Highly recombinogenic regions at seed storage protein loci on chromosome 1DS of *Aegilops tauschii*, the D-genome donor of wheat. Genetics, **155**: 361–367.

Spielmeyer, W., Huang, L., Bariana, H., Laroche, A., Gill, B.S., and Lagudah, E.S. 2000*b*. NBS–LRR sequence family is associated with leaf and stripe rust resistance on the end of homeologous chromosome group 1S of wheat. Theor. Appl. Genet. **101**: 1139–1144.

Stein, N., Feuillet, C., Wicker, T., Schlagenhauf, E., and Keller, B. 2000. Subgenome chromosome walking in wheat: A 450-kb physical contig in *Triticum monococcum* L. spans the *Lr10* resistance locus in hexaploid wheat (*Triticum aestivum* L.). Proc. Natl. Acad. Sci. U.S.A. **97**: 13 436 – 13 441.

Werner, J.E., Endo, T.R., and Gill, B.S. 1992. Toward a cytogenetically based physical map of the wheat genome. Proc. Natl. Acad. Sci. U.S.A. **89**: 11 307 – 11 311.

Wicker, T., Stein, N., Albar, L., Feuillet, C., Schlagenhauf, E., and Keller, B. 2001. Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. Plant J. **26**: 307–316.

Yang, Q., Reinhard, K., Schiltz, E., and Matern, U. 1997. Characterization and heterologous expression of hydroxycinnamoyl/benzoyl-CoA : anthranilate *N*-hydroxycinnamoyl/benzoyltransferase from elicited cell cultures of carnation, *Dianthus caryophyllus* L. Plant Mol. Biol. **35**: 777–789.

Yang, Q., Grimmig, B., and Matern, U. 1998. Anthranilate *N*-hydroxycinnamoyl/benzoyltransferase gene from carnation: rapid elicitation of transcription and promoter analysis. Plant Mol. Biol. **38**: 1201–1214.